Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors

Dellwo, Volker ; Leemann, Adrian ; Kolly, Marie-José

Abstract: Between-speaker variability of acoustically measurable speech rhythm [%V, deltaV(ln), deltaC(ln), and deltaPeak(ln)] was investigated when within-speaker variability of (a) articulation rate and (b) linguistic structural characteristics was introduced. To study (a), 12 speakers of Standard German read seven lexically identical sentences under five different intended tempo conditions (very slow, slow, normal, fast, very fast). To study (b), 16 speakers of Zurich Swiss German produced 16 spontaneous utterances each (256 in total) for which transcripts were made and then read by all speakers (4096 sentences; 16 speaker

DOI: https://doi.org/10.1121/1.4906837

Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors

Volker Dellwo[a)]
Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, CH-8032 Zurich, Switzerland

Adrian Leemann
Department of Theoretical and Applied Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge, CB3 9DA, United Kingdom

Marie-José Kolly[b)]
Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, CH-8032 Zurich, Switzerland

(Received 31 August 2013; revised 23 July 2014; accepted 5 January 2015)

Between-speaker variability of acoustically measurable speech rhythm [%*V*, Δ*V*(ln), Δ*C*(ln), and Δ*peak*(ln)] was investigated when within-speaker variability of (a) articulation rate and (b) linguistic structural characteristics was introduced. To study (a), 12 speakers of Standard German read seven lexically identical sentences under five different intended tempo conditions (very slow, slow, normal, fast, very fast). To study (b), 16 speakers of Zurich Swiss German produced 16 spontaneous utterances each (256 in total) for which transcripts were made and then read by all speakers (4096 sentences; 16 speaker × 256 sentences). Between-speaker variability was tested using analysis of variance with repeated measures on within-speaker factors. Results revealed strong and consistent between-speaker variability while within-speaker variability as a function of articulation rate and linguistic characteristics was typically not significant. It was concluded that between-speaker variability of acoustically measurable speech rhythm is strong and robust against various sources of within-speaker variability. Idiosyncratic articulatory movements were found to be the most plausible factor explaining between-speaker differences. © *2015 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4906837]

[MAH] Pages: 1513–1528

I. INTRODUCTION

Speech is highly organized in time. In the present paper we studied the degree to which suprasegmental timing patterns of speech that are assumed to be correlates of perceived speech rhythm remain constant between speakers when sources of within-speaker variability are strong. We identified possible sources of between-speaker rhythmic variability.

Why should speech rhythm vary between speakers? Speakers' voices are individual which is why listeners can typically identify speakers relatively accurately and automatic speaker recognition systems reveal high identification rates. It is well known that time-invariant characteristics of speech like voice quality and overall spectral envelope characteristics play an important role in human and automatic speaker identification (Nolan, 2002; McDougall, 2004, 2006; Dellwo *et al.*, 2007). This is based on the rationale that frequency domain parameters are to a large degree the result of individual physiological characteristics of a speaker's organs of speech.

The individual characteristics of the articulators, however, not only have a strong influence on speech frequency characteristics, a hitherto rather neglected assumption is that they also influence speech temporal organization. Speech is produced by a highly complex system of muscles, ligaments, bones, cartilages, and other biological matter forming a mechanical structure, the articulators. According to Perrier (2012), four dynamical properties are crucial in controlling an articulator, which are its mass, its damping characteristics, its stiffness, and external forces (e.g., friction) acting on it. These dynamical properties are in return the basis for observable kinematic properties, i.e., the spatial path, the velocity or the acceleration characteristics of the articulators. Given that the articulators of individuals are not equal, it must evidently be the case that their dynamic and thus their kinematic properties vary according to how fast they move, their rates of acceleration and deceleration, and the spatial dimension they occupy. This belief is supported by findings from gait research showing that temporal information derived from the movement of different anchor points (mainly along a walker's leg) during walking is highly walker-specific and that walkers can be identified based on this information fairly accurately (Loula *et al.*, 2005; Nixon, 2008). It seems conceivable that an analogous situation is true in the case of articulatory movements and that such individual movement characteristics should be encoded in the acoustic signal (Mark Nixon and Anders Eriksson, personal communication). Support for this view can be found from studies on several languages. For English, McDougall (2004, 2006) showed that temporal information derived from the dynamics of formant frequencies is speaker-idiosyncratic. Further, temporal characteristics derived from selected speech segments (consonants and vowels) have been demonstrated to vary systematically between speakers (for French,

[a)]Author to whom correspondence should be addressed. Electronic mail: volker.dellwo@uzh.ch
[b)]Also at: LIMSI-CNRS, Rue John von Neumann, Campus Universitaire d'Orsay, Bât 508, 91405 Orsay CEDEX, France.

O'Shaughnessy, 1984; for Dutch, van den Heuvel *et al.*, 1994; for Spanish, Mendoza *et al.*, 2003). Finally, Shriberg *et al.* (2005) showed for English that within-syllable temporal information (duration from the syllable onset to the nucleus or from the nucleus to the syllable offset) is speaker-idiosyncratic and that such information may be used for automatic speaker recognition purposes. Beyond a segmental or syllabic level, however, it seems further conceivable that a temporal organization might exist above the syllable, i.e., it might systematically affect the rhythmic organization of speech. Strong support for this view can be found from the paradigm of speech rhythmical measures discussed in Sec. I A.

A. Rhythmic variability in speech

Research on speech rhythm has mainly focused on language-specific rhythmic characteristics (e.g., so-called stress- and syllable-timed languages). By now, there is a wide body of evidence showing that durational characteristics of consonantal and vocalic intervals (henceforth C- and V-intervals;[1] Ramus *et al.*, 1999; Grabe and Low, 2002; Dellwo, 2006; White and Mattys, 2007) are a correlate of between-language rhythmic variability. Consonantal and vocalic durational variability is influenced by the phonology of a language (Dauer, 1983). As a means of quantification, Ramus *et al.* (1999) introduced the standard deviation of vocalic (ΔV) and consonantal (ΔC) intervals and the percentage over which speech is vocalic ($\%V$). Grabe and Low (2002) introduced the pairwise variability index (PVI), a measure of the average differences between consecutive consonantal or vocalic intervals. Variants of theses measures have been developed, such as normalizing ΔC and ΔV for speech rate variability (*V*arco*C* and *V*arco*V*, respectively, Dellwo, 2006; White and Mattys, 2007). An overview of these measures is provided in Loukina *et al.* (2011).

Whether rhythmic differences between languages exist and whether languages can be categorized according to speech rhythm is a matter of debate (White and Mattys, 2007; Dellwo, 2010; Loukina *et al.*, 2011; Arvaniti, 2012). To avoid confusion with previous studies, we continue to refer to the measures described above as "rhythm measures" even though definitions of speech rhythm are variable and the concept as such is controversial. As the measures in question calculate temporal phenomena over a period of time consisting of several words (typically a sentence), we argue that they are characterized by suprasegmental phenomena that are recurring over time. Even if these measures do not provide a comprehensive model of speech rhythm, they should certainly be strongly related to such phenomena.

The discussion about the definition of speech rhythm and its language-specific characteristics is only of secondary relevance to the present study. More important is the fact that by now there exists evidence from a number of different datasets that rhythm measurements based on vocalic and consonantal intervals can vary significantly within a language as a function of speaker (Wiget *et al.*, 2010; Yoon, 2010; Loukina, 2011; Arvaniti, 2012; Dellwo *et al.*, 2012; Leemann *et al.*, 2014). For five speakers of English, Wiget *et al.* (2010) showed that there is significant variability of $\%V$ and *V*arco*V* between speakers but not for the pairwise vocalic variability measure *n*PVI. Yoon (2010) analyzed ten speakers from the same language variety of Northern American English (Ohio variety) from the Buckeye Corpus and found similar effects in spontaneously produced speech. Earlier, but in a very similar vein, Johnson and Hollien (1984) showed that temporal information derived from the amplitude envelope and voiced and voiceless intervals of the speech signal are speaker-individual and that such information is considerably robust towards voice disguise. Speaker-specific information in the durations of voiced and voiceless intervals were also reported by Dellwo and Fourcin (2013) and Leemann *et al.* (2014). In Dellwo *et al.* (2012) and Leemann *et al.* (2014) we described the Temporal Voice Idiosyncrasy Corpus (TEVOID Corpus), and showed consistent variability of temporal patterns between 16 speakers of Zurich German. This database has been used in experiment 2 (below) where it is described in more detail.

B. Sources of rhythmic between-speaker variability

Results from previous research demonstrated that rhythmic characteristics of speech are idiosyncratic. It is possible that this might be precisely the result of idiosyncratic movement behavior of the articulators as described above. However, here we hypothesize that there are two other obvious sources that could have an influence on idiosyncratic rhythmic behavior. First, numerous studies reported that individual sentences have a large influence on speech rhythmic characteristics (e.g., Dellwo, 2010; Wiget *et al.*, 2010; Arvaniti, 2012). Ratio measures like $\%V$ as well as rate normalized or non-normalized measures of consonantal or vocalic interval variability have been shown to vary drastically and consistently between sentences. This variability can be larger in magnitude than between-language variability (Wiget *et al.*, 2010). It is thus possible that speakers create an idiosyncratic rhythm by choosing lexical items and/or morphosyntactic constructions that lead to certain rhythmic characteristics when producing speech spontaneously, for example. This seems even more likely in the light of results which show that syllable structure plays an important role within languages in that sentences characterized by predominantly phonotactically simple syllables reveal measurable rhythmic differences from their more complex peers (Prieto *et al.*, 2012). An idiosyncratic choice of words or morphosyntactic patterns containing predominantly simple or complex phonotactic characteristics could thus influence measurable speech rhythmic characteristics (henceforth: linguistic factors). Second, speech rhythm together with intonation and stress is grouped together to a phenomenon typically referred to as prosody. It seems conceivable that other prosodic factors like intonation or stress have an influence on durational aspects of speech rhythm. This view is supported by Prieto *et al.* (2012) who found that the stressing of prosodic heads or pre-final syllables leads to systematic variability in measurements of speech rhythm. It also seems feasible that the prosodic use of intonation patterns has an influence on acoustically measurable speech rhythm (certain intonational movements may require more time than others;

Kohler, 1983). So a speaker's idiosyncratic speech rhythm might in part be a result of an idiosyncratic use of, for example, stress patterns and/or intonation (henceforth: prosodic factors).

C. Aims of the present experiments

In summary, previous research provides strong evidence for speech rhythm to be speaker idiosyncratic. It seems likely that the sources for the variability between speakers are of articulatory, linguistic, and/or prosodic nature. The present paper aims at enhancing our understanding of speaker individual rhythmic characteristics with a possible application of the results for speaker identification purposes in mind. Variables for speaker recognition are powerful when their between-speaker variability is high and their within-speaker variability is low (Nolan, 2009). For this reason we tested how robust between-speaker variability of speech rhythm remained when we introduced within-speaker prosodic (experiment 1) and linguistic variability (experiment 2) was strong. By studying these two sources of variability we aimed to interpret the strength of the third source of variability, articulatory movements, which was not tested specifically in these experiments. In experiment 2 we further tested whether we can normalize for the influence of linguistic factors.

In experiment 1 we introduced within-speaker variability by studying speech containing extreme rate variability. We aimed to test whether between-speaker differences persist in cases of substantial prosodic within-speaker variability. Within-speaker linguistic variability was introduced in experiment 2 by letting speakers read sentences that they either generated themselves or that other speakers generated for them. We tested the influence of idiosyncratic linguistic (lexical and morphosyntactic choices) characteristics on speaker-individual rhythm.

II. SELECTION OF RHYTHM AND RATE MEASURES

Measures of speech rhythm can be subdivided into two categories (Tilsen and Arvaniti, 2013), measures based on (a) speech interval durations (Sec. I A) and (b) temporal characteristics of the amplitude envelope (Sec. I B). For the present study, we included measures from both domains. We selected existing measures for (a) and created a new measure for (b). Since we were dealing with variable rates (in particular, in experiment 1 where speakers were asked to vary their speech tempo) we also selected a measure of speech rate.

A. Interval-based rhythm measures

These rhythm measures can be roughly categorized into three classes: consonantal and vocalic durational ratio measures (percentage over which speech is vocalic, $\%V$), consonantal and vocalic durational variability measures (standard deviation of consonantal or vocalic interval durations, ΔC and ΔV; average durational differences between consecutive consonantal or vocalic intervals, rPVI) and rate-normalized consonantal and vocalic variability measures (coefficient of variation of consonantal or vocalic interval durations, Varco$C$ and Varco$V$; average differences between consecutive consonantal or vocalic intervals proportional to the duration of an interval pair, nPVI). As speakers vary in speech rate, we excluded non-rate normalized measures (ΔC, ΔV, rPVI). A widely applied normalization procedure for rate is the coefficient of variation (Varco$C$ and Varco$V$, respectively, Dellwo, 2006; White and Mattys, 2007). Dellwo (2009), however, demonstrated that the durations of consonantal and vocalic intervals are non-normally distributed (highly negatively skewed and a high degree of kurtosis) which is why the calculations of standard deviations or coefficients of variation are problematic as they do not represent the underlying data distributions well. Since it is possible that speakers vary systematically in the degree of skewness and kurtosis, this procedure is prone to create artifacts in obtaining between-speaker effects. To address this problem, Dellwo (2009) calculated ΔV and ΔC on durations that are expressed as logarithms to the base e. This procedure resulted in normally distributed durations of vocalic and consonantal intervals and, in addition, it normalized for speech rate variability. For the rate-normalized PVI (nPVI), Wiget *et al.* (2010) did not obtain any speaker-specific effects; hence we excluded this measure from our analysis.

B. Amplitude envelope-based rhythm measures

Other approaches to measuring speech rhythm exist that are less drawn to segmental properties (such as the measures in Sec. II A) but rather to acoustically recurring information such as amplitude beats derived from the amplitude envelope of speech. These approaches draw on the theory that syllables contain a perceptual center (p-center) for which the acoustic correlates are a complex mixture of amplitude envelope peaks, fundamental frequency movements and segmental qualities (Morton *et al.*, 1976). It has been argued that the temporal characteristics of syllabic beats are more salient in terms of the perceptual rhythmic characteristics of speech than are syllabic or segmental boundaries (Tilsen and Johnson, 2008; Tilsen and Arvaniti, 2013) and that acoustic syllabic beats of different magnitude may occur at oscillating intervals which produces a regularity in the rhythmic structure of speech (coupled oscillator models, O'Dell and Nieminen, 1999). Recent approaches on the basis of salient low frequency characteristics of the amplitude envelope of speech are used in a model based on Fourier transforming a low-pass filtered waveform (Tilsen and Johnson, 2008; Tilsen and Arvaniti, 2013).

In the present paper, we applied a measure that we first developed in Dellwo *et al.* (2012) which monitors the variability of intervals between syllabic beats by calculating the standard deviation of interval durations between syllabic amplitude peak points (inter-peak intervals). Even though syllabic amplitude peaks are not the only correlate of a perceptual syllabic center (Howell, 1988) we found that it is an approximation that might be particularly suitable from a production point of view. Since amplitude peak points most

likely occur at a maximum mouth aperture or a maximum of vocal fold activity, it seems conceivable that these points also correlate with turning points in articulation. Therefore, if speakers' individual articulatory movements are responsible for a speaker's idiosyncratic rhythm, then we might expect the durational organization between amplitude peaks to reflect this. Inter-peak intervals were defined as the interval between the amplitude maximum in the amplitude envelope of a vocalic interval (as the nucleus of the syllable) and the amplitude maximum in the amplitude envelope of the following vocalic interval, hence, this method excluded syllabic consonants. The first inter-peak interval in an utterance was always the interval between the first and second vocalic amplitude peak, the last interval between the pre-final and final vocalic interval amplitude peak. This means that the signal parts from the utterance onset to the first vocalic peak as well as from the last vocalic peak to the utterance offset were not part of the analysis. The amplitude envelope of a signal was extracted by half-wave rectifying the signal and then low-pass filtering it at 10 Hz. The identification of inter-peak intervals was performed with PRAAT (Boersma and Weenink, 2013) using the script durationTierCreator.praat (Dellwo, 2013). We calculated the standard deviation of the inter-peak interval durations ($\Delta peak$) for each sentence. As the frequency distributions of inter-peak durations showed a similar degree of skewness and kurtosis as consonantal and vocalic intervals, we also calculated $\Delta peak$ based on log transformations of the raw durations [$\Delta peak$(ln)].

C. Speech rate measure

We used the number of consonantal or vocalic intervals per second (rateCV) as a correlate of articulation rate since this is based on the same intervals that we used for the rhythm measures described above. Since there is typically a vocalic interval at each syllabic nucleus (in the database for experiment 2: 99.1% of syllables contain a vocalic nucleus) the number of consonantal and vocalic intervals is close to exactly twice as high as the number of syllables (a unit that is possibly more commonly used as a correlate of articulation rate).

D. Summary

We have chosen five temporal measures, one rate measure (rateCV), one durational consonantal-vocalic ratio measure ($\%V$), and three interval variability measures, two based on consonantal and vocalic intervals [ΔV(ln), ΔC(ln)] and one based on inter-peak intervals [$\Delta peak$(ln)]. The measures were calculated as follows:

$$\%V = \frac{\left(\sum_{i=1}^{N_V} V_i\right) 100}{\sum_{i=1}^{N_C} C_i + \sum_{i=1}^{N_V} V_i} \qquad \begin{aligned} N_V &= \text{number of } V\text{-intervals in sentence},\\ N_C &= \text{number of } C\text{-intervals in sentence},\\ V_i &= \text{duration of the } i\text{th } V\text{-interval},\\ C_i &= \text{duration of the } i\text{th } C\text{-interval}. \end{aligned} \tag{1}$$

ΔV(ln), ΔC(ln), and $\Delta peak$(ln) were calculated according to the following equation:

$$\Delta \text{Int ln} = \sqrt{\frac{N_{\text{Int}} \sum_{i=1}^{N_{\text{Int}}} (\ln \text{Int}_i)^2 - \left(\sum_{i=1}^{N_{\text{Int}}} (\ln \text{Int}_i)\right)^2}{N_{\text{Int}}(N_{\text{Int}} - 1)}} \qquad \begin{aligned} \text{Int} &= \text{interval under observation}\\ &\quad (\text{either } V, C, \text{ or inter-peak}),\\ N_{\text{Int}} &= \text{number of respective intervals in sentence},\\ \text{Int}_i &= \text{duration of the } i\text{th interval}. \end{aligned} \tag{2}$$

The acoustic measure of speech rate was calculated as follows:

$$\text{rate } CV = N_{CV}/d, \tag{3}$$

where N_{CV} is the number of C- or V-intervals in the sentence and d is the duration of the sentence in s (excluding pauses).

III. EXPERIMENT 1: THE INFLUENCE OF WITHIN-SPEAKER RATE VARIABILITY ON BETWEEN-SPEAKER RHYTHMIC DIFFERENCES

A. Introduction

Between-speaker rhythmic variability of speech rhythm was studied when within-speaker articulation rate variability was high. First evidence that speakers' rhythmic signature remains constant when prosodic variability increases has been demonstrated in our previous work (Leemann *et al.*, 2014) where we created within-speaker variability by letting speakers produce speech under varying speaking styles (spontaneous and read speech). Because of the strong variability of acoustic rhythm as a factor of sentence (Wiget *et al.*, 2010; Dellwo, 2010), we elicited read speech based on transcripts of sentences previously spontaneously produced by the speakers. In the present experiment we enforced the within-speaker variability to a higher degree. In the present experiment we studied speech of German speakers from the BonnTempo Corpus (Dellwo *et al.*, 2004; Dellwo, 2010). In this corpus, prosodic variability was introduced by asking speakers to read speech in a normal, slow, very slow, fast,

and fastest possible intended tempo. Such changes in the intended tempo not only result in faster and slower measurable acoustic correlates of speech rate (e.g., the number of syllables per second) but create substantial variability in the quality of intonation contours, number of intonation phrases/prosodic chunking, coarticulatory phenomena, segmental reduction phenomena, syllabic reduction phenomena, phonological elisions, etc. (Kohler, 1983; Caspers and van Heuven, 1995; Fougeron and Jun, 1998; Trouvain and Grice, 1999). A drastic increase of acoustically measurable rate as a function of intended tempo (from very slow to very fast reading) has been shown for this data (Dellwo and Wagner, 2003; Dellwo, 2010). Here we tested in which way speaker-specific durational characteristics of intervals such as consonantal, vocalic and interpeak intervals would be affected by this variability. We studied the following effects.

(a) The effects of rate variability on structural changes in speech. We studied the variability of pausing and the relative frequency of consonantal and vocalic intervals between speakers and tempo conditions.
(b) Within- and between-speaker variability of speech rhythm. We ran analyses of variances (ANOVAs) with each rhythm measure as a dependent variable and repeated measures on the within-speaker factor. We argue that rhythm measures contain particularly strong speaker-specific information when a main effect of speaker can be obtained in the absence of a main effect of any of the within-speaker factors (tempo, linguistic variability). However, a main effect of speaker in the presence of main effects of within-speaker factors may also provide us with useful information about between-speaker rhythmic variability as long as there is no interaction between the two factors. An interaction would imply that individual speakers behave differently at different levels of within-speaker variability, a situation where between-speaker effects are hard to interpret.
(c) Effects of between-speaker structural variability on between speaker rhythm. We tested the difference of pausing and the number of consonantal and vocalic intervals realized in the read speech between speakers.
(d) Effects of sentence. We tested the rhythmic variability between sentences and whether this variability can be explained by sentence structural differences.

B. Method

1. Speakers

12 speakers (5 male, 7 female) of Standard German [mean age: 30.3 years; standard deviation: 6.6 years; age range: 24–48] from the BonnTempo database (Dellwo *et al.*, 2004; Dellwo, 2010) were analyzed. All speakers were standard German speakers from different regions in the central west of Germany and revealed few accentual features of their place of origin.

2. Recording procedure

Each speaker read a text consisting of 76 phonological syllables (see Appendix A) from a German novel by Bernhard Schlink (*Selbs Betrug*). Speakers were recorded several times reading the text with instructions given in the following order: (a) read the text (normal reading condition), (b) read the text more slowly than in the first recording (slow reading condition), (c) read the text even more slowly than in previous recording (very slow reading condition), (d) read the text faster than normal (fast reading condition), and (e) read the text as fast as possible (very fast reading condition).

For recordings (a)–(d), speakers typically needed one attempt. In case speakers produced a reading mistake they were asked to start reading again from the beginning of the sentence where the mistake occurred. For recording (f), speakers had as many attempts as required for them to reach a tempo they considered highest for them. Speakers typically conducted about five attempts to reach their highest tempo (lowest number of attempts: three, highest: eight). All speakers were recorded in an anechoic chamber at the former Institute for Communication Research and Phonetics of Bonn University. Recordings were made directly on PC using a large diaphragm condenser microphone (sampling rate: 44 100 samples/s; quantization: 16 bit).

3. Data editing and segmentation

Vocalic and consonantal intervals were labeled manually by Dellwo and Wagner (Dellwo *et al.*, 2004). Vocalic intervals consisted of any number of consecutive vocalic segments between the offset of the preceding and the onset of the following consonant. Consonantal intervals were labeled analogously. Silences longer than 50 ms were labeled as a pause. In cases of laryngealization, the last glottal transient of the laryngealization was chosen. Laryngealization, however, was weak in all speakers (see Sec. III A 1). In cases where voiced consonants preceded or followed a vocalic interval, the first and last glottal pulse of the interval was determined by identifying points of rapid change in spectral dynamics in the spectrogram.

4. Data analysis and statistics

The reading text was subdivided into seven syntactic intervals that corresponded either to a syntactic main- or sub-clause (intervals are indicated by vertical lines in Appendix A). For simplicity, these intervals are henceforth referred to as "sentences" even though from a grammatical point of view this might be debatable. Each rhythm measure was calculated for each acoustic signal corresponding to a sentence ($N = 420$, 7 sentences $\times$ 5 tempo versions $\times$ 12 speakers). All calculations of rhythm and rate measures were made with a PRAAT-script durationAnalyzer.praat (Dellwo, 2013). The effects of speaker and tempo were tested by ANOVA analysis with repeated measures on speaker and/or tempo where applicable (using R statistics software). Distributions of the dependent variable data were tested visually using frequency histograms. All variables were found to be unimodally distributed resembling a Gaussian

TABLE I. Number of consonantal and vocalic intervals (*C*, *V*) as well as pauses (rows) for the five different tempo conditions (columns). The brackets behind the pause numbers indicates the number of pauses between sentences of the text (first number) and the number of pauses within sentences (second number).

Tempo	1	2	3	4	5
C-intervals	851	820	797	777	728
V-intervals	803	776	764	758	726
All pauses	136 (70, 66)	114 (70, 44)	71 (62, 9)	50 (43, 7)	11 (7, 4)

bell shape. Correlations were carried out between all dependent variable pairs to test the degree to which variables might explain each other.

C. Results

Cross-plots showed that measures poorly predicted each other (Pearson's r ranged between -0.25 and 0.17). A visible inspection of cross-plots of each possible dependent measure pair confirmed this result.

1. Structural variability between tempo versions and speakers

Table I shows that the number of *C*- and *V*-intervals as well as the number of pauses decreased with the tempo condition. This is particularly true for the decrease of pauses by over 90% from $N = 136$ in the slowest version to $N = 11$ in the fastest version. The number of *C*- and *V*-intervals also decreased, but at much smaller numbers. It is apparent, however, that the loss of *C*-intervals with an increase in tempo (from $N = 851$ to 728) was stronger than the loss of *V*-intervals (from $N = 803$ to 726). A χ^2 test revealed that the relative proportions of the number *V* and *C*-intervals differed significantly between tempo versions ($\chi^2[8] = 109.72$; $p < 0.001$). This is strong evidence for a structural and prosodic reorganization of speech from the slow to the fast version. With a high number of pauses in the slowest tempo version there is a much higher number of intonation phrases leading to strongly variable intonation contours and a higher number of phrase final lengthening cases (Vaissière, 1983, p. 57).

Structural differences, i.e., variability in the ratio of *C*- and *V*-intervals as well as in pausing behavior, between speakers are provided in Table II. The largest difference obtained for the number of *C*-intervals was $N = 45$ (between speakers 5 and 9; ratio = 1:1.14), which is a difference of about 13% between these two speakers. The largest *V*-interval difference was $N = 36$ (again between speakers 5 and 9 = 1.12). The largest pause difference was $N = 29$ (between speakers 4 and 9; ratio = 1:2.9). This means that the maximum speaker differences for the number of *C*- and *V*-intervals used were not as drastic being only about 1.13 times more but the number of pauses could vary drastically with speaker 9 creating about three times more pauses than speaker 4, for example. The numeric differences between speakers were found to be significant ($\chi^2[22] = 38.75$; $p = 0.016$). To test how the individual interval types (*C*-, *V*-intervals and pauses) varied between the tempo conditions we carried out one χ^2 test for each interval type (Bonferroni corrected α: 0.033 [0.05/3]). Results revealed that both for *C*- and *V*-intervals the differences were non-significant (*C*: $\chi^2[11] = 5.65$, $p = 0.9$; *V*: $\chi^2[11] = 4.76$, $p = 0.94$) but for pauses it was highly significant ($\chi^2[11] = 39.19$, $p < 0.001$). This means that the main structural differences between speakers were in the number of pauses they applied. It must also be noted that the reduction of consonantal and vocalic intervals is to the largest degree the result of the loss of pauses; a loss of a pause between equal interval types creates one out of previously two intervals.

2. Rate and rhythm variability between speakers and tempo versions

Figure 1 contains boxplots of rate*CV* and all rhythm measures under observation for speakers and intended tempo. For each variable a two-factor ANOVA (speaker * tempo) with repeated measures on speaker and tempo [R code: aov(dependent ~ speaker * tempo + error(sentence/(speaker*tempo)), data = data)] was carried out. *F*-values with their corresponding probability values for the main effects (speaker and tempo) as well as the interactions (speaker:tempo) are reported in Table III. Because we carried out multiple tests on the same dataset we tested at a conservative α-level of 0.01.

a. rateCV. Speech rate (rate*CV*) increased strongly from the slowest to the fastest intended tempo category (Fig. 1) and there were strong differences between some of the speakers. Both effects were significant (Table III, row 1), however, the effects were not readily interpretable as their interaction was significant as well. For this reason, simple effect tests for both speaker and tempo were carried out (Bonferroni adjusted α-levels; speaker: $\alpha = 0.0008$ [0.01/12], tempo: $\alpha = 0.002$ [0.01/5]). Simple effects were tested with a one-factor ANOVA with repeated measures on the respective simple effect [speaker or tempo; R code: aov(dependent ~ factor + error(sentence/factor), data = subset.data)]. Simple effects of speaker revealed that at each tempo level, the main effect of speaker was significant. *F*[11,83] was highest for the very slow version (51.4) and decreased with an

TABLE II. Number of consonantal and vocalic intervals (*C*, *V*) as well as pauses (rows) for the 12 different speakers.

Speaker	1	2	3	4	5	6	7	8	9	10	11	12
C-intervals	343	340	329	329	357	315	324	343	312	319	333	329
V-intervals	331	329	317	311	341	305	314	330	305	307	323	314
Pause	35	20	21	49	41	41	34	22	17	39	23	40

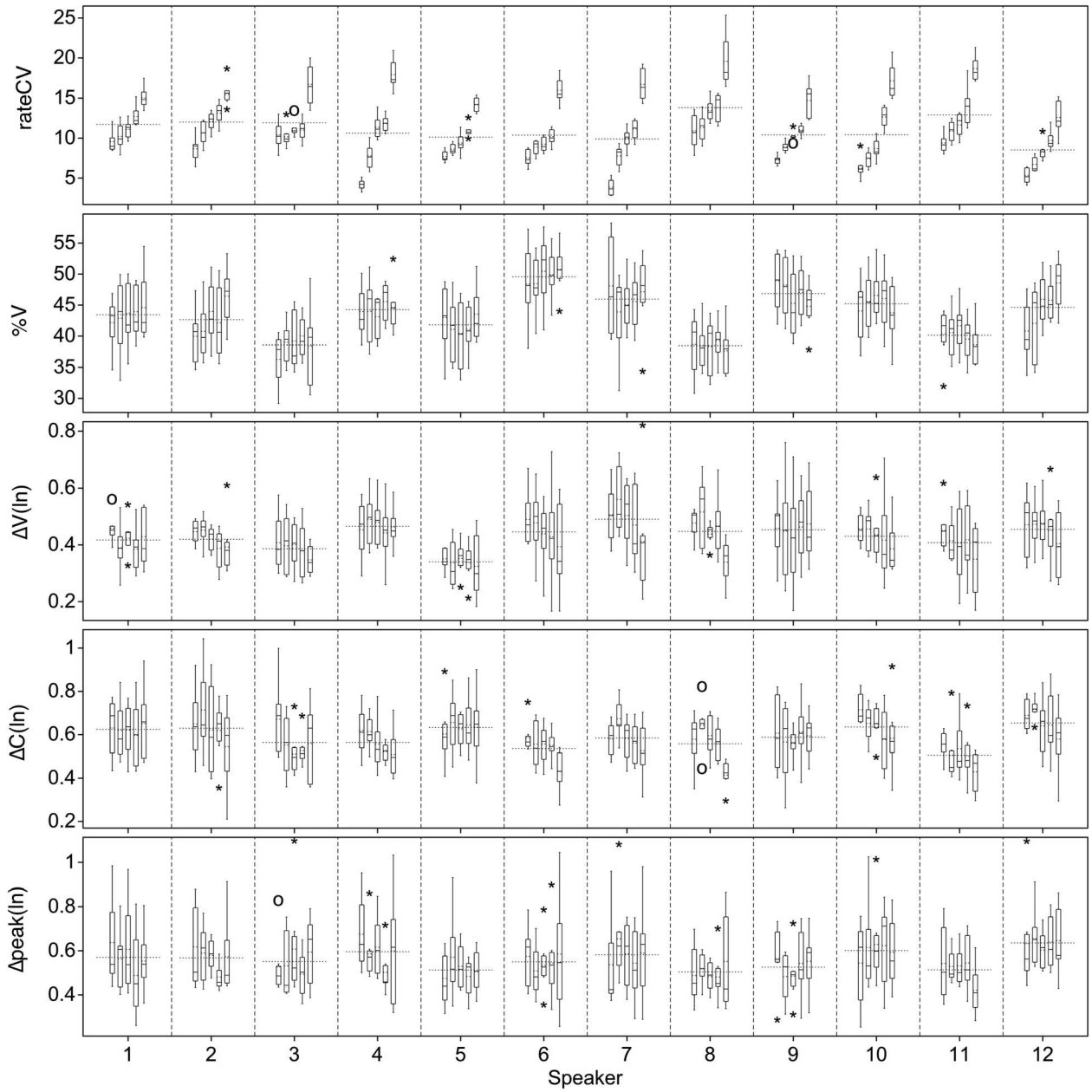

FIG. 1. Boxplots showing the distributions of the variables (a) rate*CV*, (b) *%V*, (c) Δ*V*(ln), (d) Δ*C*(ln), and (e) Δ*peak*(ln) as a function of speaker (left) and of intended tempo (right).

TABLE III. *F*-values and probability values (in brackets) for a two-factor ANOVA with repeated measures on speaker and tempo for each variable (rows) under investigation. Significant effects are highlighted in bold ($N = 420$; $\alpha = 0.01$, adjustment for multiple tests on the same data set; DOF = degrees of freedom).

Dependent	Main effect speaker (DOF: 11, 264)	Main effect tempo (DOF: 4, 264)	Interaction (speaker ∗ tempo) (DOF: 44, 264)
rate*CV*	**27.51 (<0.001)**	**404.3 (<0.001)**	**8.93 (<0.001)**
%V	**22.94 (<0.001)**	1.26 (0.31)	**3.13 (<0.001)**
Δ*V*(ln)	**3.08 (0.002)**	3.15 (0.033)	1.02 (0.45)
Δ*C*(ln)	**4.44 (<0.001)**	**4.73 (0.006)**	1.28 (0.13)
Δpeak(ln)	**3.99 (<0.001)**	0.65 (0.63)	1.22 (0.174)

increase in tempo version (slow: 22.7, normal: 22.9, fast: 11.8, fastest possible: 7.7; all $p < 0.0008$). Simple effects of tempo showed highly significant rate*CV* differences between tempo levels for each speaker [F[4,34] between 29.7 (lowest) and 180.4 (highest)]. The results revealed that speech rate varied strongly within each speaker between the slowest and the fastest tempo category. Figure 1 shows that some speakers (e.g., 3, 5, and 6) have a less strong increase in rate*CV* in particular between the very slow and the fast tempo. We ran *post hoc* comparisons for the tempo categories for each speaker (results not shown) and found that rate*CV* was always significantly different between the very slow, normal, and fastest possible tempo categories. Concerning the adjacent categories

(very slow-slow, slow-normal, normal-fast, fast-fastest possible) some speakers revealed significant differences, others did not. It is very likely that this variability contributed to the interaction. Given that there were only seven data points per speaker in each tempo condition *post hoc* effects between the adjacent conditions are difficult to obtain. We can conclude, however, that rate variability from very slow to very fast rates for each speaker was successfully obtained in our data.

b. %V. Figure 1 suggests that differences between speakers can be high and that the differences between the tempo versions are low. Inferentially this impression was confirmed by a highly significant main effect of speaker and no significant effect of tempo (Table III, row 2). As with rate*CV*, there is a highly significant interaction. Simple effects were examined to interpret the main effects. At each tempo level, simple effects of speaker were highly significant ($p < 0.0008$). For each speaker the effects of tempo were not significant apart from speakers 2, 9, and 12, however, there was no unified direction of the effect. It is also apparent from Fig. 1. that the rate variability within a speaker did not have a systematic influence on the *%V* variability. For example, speaker 4 who shows the strongest differences in rate*CV* between the very slow and the fastest possible tempo versions shows very little *%V* variability across the tempi. It can be concluded that between-speaker effects are strong and present throughout the data while within-speaker differences of *%V* are typically not obtainable and if they occur, they do so in random directions.

c. Interval variability measures. For the interval variability measures ΔV(ln), ΔC(ln), and $\Delta peak$(ln) (Fig. 1), the variability between speakers seems less strong in magnitude [in particular, for $\Delta peak$(ln)] than for *%V*. All between-speaker effects, however, are highly significant (Table III, rows 3–5). Interactions between the main effects are not present which means that both main effects are interpretable. For both ΔV(ln) and ΔC(ln) a slight decrease is visible at higher tempo (Fig. 1). This effect is only significant in the case of ΔC(ln).

d. Post hoc comparison. Concerning the effects of speaker, it is evident from Fig. 1 that some speakers vary strongly and consistently for any of the variables but others also reveal very similar values. To quantify the number of differences between speakers, *post hoc* analyses were performed using Bonferroni adjusted pairwise *t*-tests [*R* function: pairwise.tst(data$dependent, data$speaker, p.adj = "bonferroni")]. For rate*CV*, 13 of the 66 (20%) possible paired comparisons are significant ($p < 0.05$). For *%V*: 50% (33/66), ΔV(ln): 10.6% (7/66), ΔC(ln): 9% (6/66), and $\Delta peak$(ln): 1.5% (1/66). This means that the highest number of significant between-speaker comparisons can be obtained with *%V*. Even though the main effects are significant in all cases, only a few speakers significantly vary from each other *post hoc* in ΔV(ln) and ΔC(ln), and only one speaker contrast is significant in case of $\Delta peak$(ln).

3. Effect of structural differences between speakers on rhythmic variability

To test whether the between-speaker structural differences had an influence on their rate and rhythm scores we correlated the between speaker structural differences with the average rate (rate*CV*) and rhythm scores [*%V*, ΔV(ln), ΔC(ln), $\Delta peak$(ln)]. Since we carried out multiple tests on the same dataset we tested at a conservative α level of 0.01. None of the correlations were significant; an observation of cross-plots of all variable pairs supported this result (descriptive and inferential data not shown here). These results corroborate the point that between-speaker differences in rate and rhythm characteristics are not a result of structural differences between speakers.

4. Effects of sentence

The influence of sentence on measures of speech rhythm was strong and consistent as can be seen in the boxplots in Fig. 2. Descriptively, however, there was little influence of sentence on rate*CV*. For all variables, however, one-way ANOVAs with repeated measures on sentence [R-code: aov(dependent ~ sentence + error(speaker)] revealed that sentence effects were highly significant [rate*CV*: $F[6,402] = 3.43$, $p = 0.003$, *%V*: 83.99, $p < 0.001$, ΔV(ln): 22.77, $p < 0.001$, ΔC(ln): 20.37, $p < 0.001$, $\Delta peak$(ln): 53.85, $p < 0.001$).

The number of *C*- and *V*-intervals varied strongly between sentences (*C*-intervals from sentences 1–7: 384, 526, 654, 399, 656, 442, 424; *V*-intervals: 376, 531, 710, 458, 603, 499, 487; pause: 21, 19, 28, 16, 38, 5, 3). A χ^2 test revealed that this variability between sentences was highly significant ($\chi^2[12] = 45.32$, $p < 0.001$). This highly significant effect could also be replicated for each of the interval types (Bonferroni corrected $\alpha = 0.003$ [0.01/3]; *C*: $\chi^2[6] = 163.7$, $p < 0.0003$; *V*: $\chi^2[6] = 132.1$, $p < 0.0003$; pause: $\chi^2[6] = 48.8$, $p < 0.0003$). These results strongly support the view that structural characteristics like the number of *C*- or *V*-intervals play a role for between sentence rhythmic differences. To study this further we correlated the interval types with the average rate and rhythm scores for each sentence. The correlation was strong between *%V* and the number of *C*-intervals ($r[6] = 0.76$) or the number of *V*-intervals ($r[6] = 0.68$) but insignificant at an α level of 0.01.

D. Discussion

Our results revealed significant variability in all tested variables of acoustically measurable speech rhythm between speakers when speech rate varied strongly within speakers. Moreover, the variability of rhythm measures as a function of tempo can be interpreted as ranging from low to non-existent. There was also a very strong variability of all measures of speech rhythm as a function of sentence, which reveals an important characteristic about the variables under observation. The sentences that have been chosen (see Appendix A) can either be grammatical main- or sub-clauses, which means that prosodic characteristics vary between them. In addition, by varying speech rate, the sentences underwent a high variability in their prosodic realization. Our structural analysis

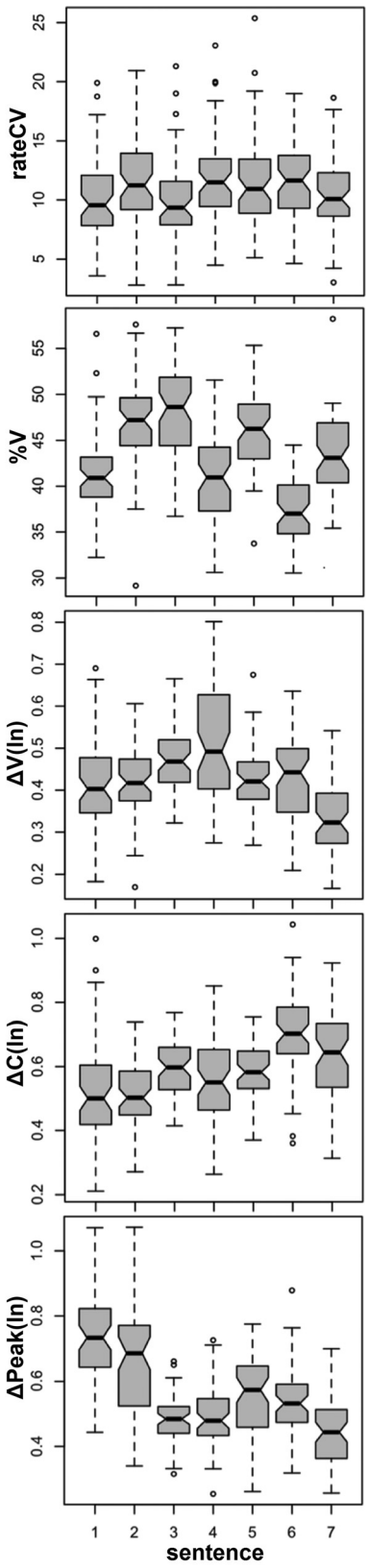

FIG. 2. Boxplots showing the distributions of each of the variables under investigation [rate*CV*, *%V*, Δ*V*(ln), Δ*C*(ln), Δpeak(ln)] for the seven different sentences.

revealed that in the slow and very slow tempo conditions (1 and 2) the discourse was interrupted by a larger number of pauses which often created two or three intonation phrases within one syntactic constituent [e.g., "*Am nächsten Tag fuhr ich nach Husum*" is realized as "*Am nächsten Tag*" (silence) "*fuhr ich nach Husum*"]. In spite of this strong prosodic variability, the differences in rhythm between the sentences remain consistent. We therefore conclude that the linguistic constituents of a sentence are probably the primary cause in rhythmic variability. This is in line with the findings by Wiget *et al.* (2010) and Prieto *et al.* (2012) (the latter showed that the phonological and phonotactic complexity of syllables strongly contribute to rhythmic variability; see discussion in Sec. I A). We showed that there was no significant variability in the number of linguistic constituents between speakers. In this experiment, speakers produced read speech, which means they did not have a choice about which linguistic constituents they could choose. In experiment II we tested whether the free choice of words and sentence structures could have an influence on between-speaker rhythmic variability by analyzing spontaneously-produced speech. A further reason for the low structural variability may also lie in the type of structural information investigated. In the BonnTempo corpus there were only syllabic and *C*- and *V*-interval boundaries available, so no further details about the internal structure of these intervals were available. It is possible that there are stronger differences between speakers in terms of the syllabic complexity they produced. In experiment II we studied this question on a larger database in which the internal structure of syllables was available (TEVOID Corpus).

What could be the reasons for rhythm measures to show consistent between-speaker variability? In the Introduction we hypothesized three possible factors: articulatory, linguistic, and prosodic individualities of the speaker. Based on the results of the present experiment it seems feasible to exclude sentence structural idiosyncrasies from responsibility for the observed between-speaker rhythmic differences as each speaker produced identical material (read speech). However, given the strong between-sentence effects discussed above it seems possible that when speakers are free to reveal their individual choice of lexical items and morphosyntactic patterns (as in spontaneous speech), this choice may contribute strongly to their rhythmic signature (this hypothesis was tested in experiment 2).

An alternative explanation for the between-speaker rhythmic variability might be that speakers maintained individual prosodic realizations of the sentences (e.g., stress-patterning or intonation), which might influence individual suprasegmental temporal characteristics in their speech. Given the finding, however, that the sentence effect remained consistent despite the strong variability of prosodic characteristics between the tempo versions, it does not seem likely that there are prosodic characteristics between speakers that could potentially explain such effects. For example, if a speaker-individual stress pattern would be the driving factor for the speaker's idiosyncratic rhythm scores, the speaker would have to maintain this characteristic under different tempo versions and different prosodic chunking. Given the results presented above, we do not find this explanation plausible. In summary, since both sentence and prosodic variability are not very plausible explanatory factors, it seems likely that idiosyncratic articulatory

movements contribute most strongly to the between-speaker rhythmic variability.

IV. EXPERIMENT 2: THE INFLUENCE OF WITHIN-SPEAKER SENTENCE VARIABILITY ON BETWEEN-SPEAKER RHYTHMIC DIFFERENCES

A. Introduction

In the present experiment we studied within-speaker linguistic variability in the TEVOID Corpus (Dellwo *et al.*, 2012). In Leemann *et al.* (2014) we showed with this dataset that speakers vary in suprasegmental temporal characteristics in a larger dataset of $N = 4096$ (16 speakers $\times$ 256 sentences) and that within-speaker variability of speaking style (spontaneous and read speech) did not have an effect on between-speaker rhythmic variability. To test this we compared rhythm scores of the 16 spontaneously produced sentences by each speaker with their read peers. Spontaneous speech can be very variable in terms of prosody compared to read speech (Lieberman *et al.*, 1985; Howell and Kadi-Hanifi, 1991). Read speech reveals no individual choice in sentence structural characteristics (choice of lexical and morphosyntactic patterns). Since sentence structural characteristics of an utterance have a high influence on rhythmic variability (experiment 1, Dellwo, 2010; Wiget *et al.*, 2010) we considered it to be essential to have the same sentences produced under both spontaneous and read speech to be able to compare like with like. To meet this constraint, we recorded 16 speakers producing spontaneous speech in interviews. We then made transcripts of 16 selected sentences (see sentence list in Appendix B) from the interview and asked speakers to read them. Each speaker read both their own previously spontaneously produced sentences as well as the transcripts of the sentences of all other speakers (256 sentences in total, 16 speakers $\times$ 16 sentences). With this design we tested the following effects.

(a) The effects of linguistic structural characteristics on between-sentence speech rhythm: This was tested by comparing the complexity of consonantal and vocalic intervals across the sentences.
(b) The variation of consonantal and vocalic complexity between speakers in read and spontaneous speech: This was tested looking at counts of intervals of varying complexity between speakers for both read and spontaneous speech.
(c) The influence of linguistic structural idiosyncrasies on between-speaker rhythmic variability: For each speaker (X), we selected a set of utterances for which the sentence structures were generated by speaker X and compared them to a set of utterances for which the sentence structures were generated by a variety of speakers (excluding X).
(d) The influence of sentence normalization procedures on between-speaker rhythm effects: Between speaker effects were calculated with and without sentence variability normalization which was performed by calculating z-scores for each sentence mean and standard deviation.

Our assumption was that if we obtained variability in the structural complexity of sentences in (a), then this variability might be present in the spontaneous speech between speakers in (b) but not in the read speech (in read speech, speakers have no choice about the complexity of consonantal and vocalic clusters). Should variability exist between speakers then sentences originally produced spontaneously by a speaker might show differences in their rhythm scores from sentences originally produced by their peers (c).

B. Method

1. Speakers

16 speakers of the Zurich variety of Swiss German were recorded (8 m, 8 f), mean age: 27 years, standard deviation 3.6 years, age range: 20–33. Speakers were either students or acquaintances of this group. Speakers were screened for their regional variety (Zurich dialect) prior to the recording by trained phoneticians (second and third author).

2. Speech material and recording procedure

An interview that lasted around 45 min was carried out with each speaker (i.e., 16 interviews). Speakers were asked about their last holidays, their fields of study, and their plans after graduation. The topics were selected such that speakers felt comfortable and could talk freely, fluently and without inhibitions. Interviews were all carried out by the same interviewers (second and third authors) in Swiss German. Both interviewers spoke with a Western Swiss German dialect. From each interview, 16 sentences were extracted resulting in 256 sentences in total (16 speakers $\times$ 16 sentences). The criteria for sentence extraction were that utterances had to be grammatically coherent without major interruptions, hesitations and pauses. We looked for sentences of about 15 syllables in length (even though this number sometimes varied considerably; see Appendix B). Of all possible candidates we randomly selected 16. Orthographic transcripts in Zurich German were made of all 256 sentences. About four weeks after the interviews, the 16 speakers were re-invited individually for a reading task in which each speaker read the 256 transcribed sentences from the interview resulting in 4096 read sentences (16 speakers $\times$ 256 sentences). As reading skills of dialect transcripts varied between the speakers—there are no formal criteria for writing in Swiss German—they were given the transcript of the sentences a few days prior to the recordings and were asked to practice them well for at least one hour to be able to read them fluently. With the self-rehearsal all speakers were able to read the sentence list fluently. Speakers received 30 CHF per hour for the interview, the reading task, and the preparation of the reading task. Both the interviews and the reading task were recorded in a sound-attenuated recording room in the Phonetics Laboratory at Zurich University. Recordings were made directly to a Mac Pro with a transducer microphone (Neumann STH) using ProTools (sampling frequency of 44 100 samples/s; 16-bit quantization).

3. Data processing

All 16 spontaneous sentences of each speaker were annotated manually with a phonetic transcription. The annotations were done in PRAAT (Boersma and Weenink, 2013) using the annotation function. From the annotated files (PRAAT TEXTGRIDS) of spontaneous speech, new files were produced automatically that matched the total duration of each respective sentence of the read speech. All 4096 automatically produced TEXTGRIDS were manually corrected by A.L. and M.-J.K. Manual correction meant adjusting the segmental boundaries and deleting, adding or modifying segments in cases where speakers deviated from the segmental content of the spontaneous version. The phonetic data labeling was automatically processed into consonantal and vocalic intervals using PRAAT scripts. Durational analysis of intervals was performed using durationAnalyzer.praat (see experiment 1). One value per sentence was calculated for each rhythm measure; *z*-score values were calculated by sentence.

C. Results

1. The effect of sentence structural characteristics on sentence rhythm scores

Between sentence variability was measured by analyzing the number of structurally different consonantal and vocalic intervals. Consonantal intervals consisted of types reaching from one to seven consonants and vocalic intervals from one to three vowels. The distribution of these intervals were *c*: 29 262, *cc*: 16 295, *ccc*: 4453, *cccc*: 890, *ccccc*: 145, *cccccc*: 21, *ccccccc*: 1, *v*: 48 051, *vv*: 2356, *vvv*: 86; total $N = 101\,560$. This data shows that the most common type of intervals are *V*, *C*, and *CC* intervals, making up about 92% of intervals in the database. As the less frequent interval types only occurred very sporadically across the 256 sentences we excluded them from the analysis (consonantal intervals with more than four consonantal segments and vocalic intervals with three segments). We studied stacked bar-plots showing the number of different intervals for each sentence (not shown). There was a large variability between sentences in the number of different intervals used. A χ^2 test revealed that this variability was highly significant ($\chi^2[1275] = 11\,693.51$, $p < 0.001$). We correlated the number of *V*-intervals and the number of *C*-intervals in a sentence with the rhythm score for that sentence. Since the number of items was rather large we used the rule $|r| >= 2/\sqrt{n}$ to determine when a correlation was indicating a relationship. This formula returns an absolute *r*-threshold of 0.125 $(2/\sqrt{256})$. Correlation results (Table IV) revealed that a negative correlation between *%V* and the consonantal intervals increases with a complexity in consonantal intervals. The existence of double vowel intervals also show a higher *%V*. A somehow opposite case seems to be present for ΔC(ln), where *ccc* and *cccc* intervals lead to an increase to consonantal durational variability. Double *V*-intervals also have an effect on ΔV(ln) and $\Delta peak$(ln). In both cases intervals made up of two vocalic segments lead to a higher vocalic durational variability. Double *V*-intervals also have a rather strong influence on rate*CV*. The lack of consonants between vowels leads to a higher number of *V*-intervals produced per second. In summary, the results from this section reveal that (a) the interval complexity varies between sentences and (b) that this variability has an influence on rhythm scores.

TABLE IV. *r*-values for the correlations between the rhythm and rate variables (rows) and the number of interval types (columns) for each sentence.

	c	*cc*	*ccc*	*cccc*	*v*	*vv*
rate*CV*	0.054	0.164	0.056	0.033	0.199	−0.516
%V	0.138	−0.221	−0.299	−0.303	−0.083	0.329
ΔV(ln)	0.024	−0.012	−0.011	−0.088	−0.079	0.493
ΔC(ln)	−0.073	−0.094	0.218	0.233	−0.070	0.008
$\Delta peak$(ln)	0.113	0.007	0.111	0.060	0.033	0.476

2. Structural differences between speakers in spontaneous and read speech

Figure 3 shows the relative frequency of the most frequent interval types between the sixteen speakers of the TEVOID corpus for spontaneous speech (left) and for read speech (right). The figure reveals that the frequencies are more equal between speakers in read speech than they are in spontaneous speech. While in spontaneous speech the frequencies of vocalic interval types (*v* and *vv*) do not vary much either between speakers, some variability can be observed for the consonantal types (*c*, *cc*, *ccc*, *cccc*). This means that speakers varied in their structural interval complexity when producing utterances for which they chose the wording themselves (spontaneous speech). This variability in interval complexity between speakers might have a direct influence on the speech temporal characteristics examined. In Sec. IV C 3 we tested whether such individual variability in segmental complexity can lead to between-speaker rhythmic variability.

3. Influence of sentence on between speaker differences

To test whether speakers' choice of sentences contributes to their individuality, for each speaker we selected the 16 read utterances that they previously produced spontaneously (henceforth: own set) and 16 read utterances, based on a randomly selected sentence from each speaker (no doublets; henceforth: mixed set). Since this choice inevitably included two utterances based on the same sentences for each speaker (the sentence that the speaker previously produced spontaneously), we excluded this sentence from the data to have 15 read sentences in the mixed set. We referred to this factor as sentence origin (sentences originated from the speaker as opposed to sentences originated from different speakers; $N = 496$, 16 speakers × 31 sentences). We carried out a two-factor mixed design ANOVA (speaker and sentence origin) with repeated measures on sentence origin [R-code: aov(dependent ∼ speaker ∗ speaking style + error (sentence/speaking style))]. The adjusted α of 0.01 was divided by 2 (0.005) since we tested another subset of the data. Table V reveals no significant interaction for all rhythm measures and in no case did we find an effect of

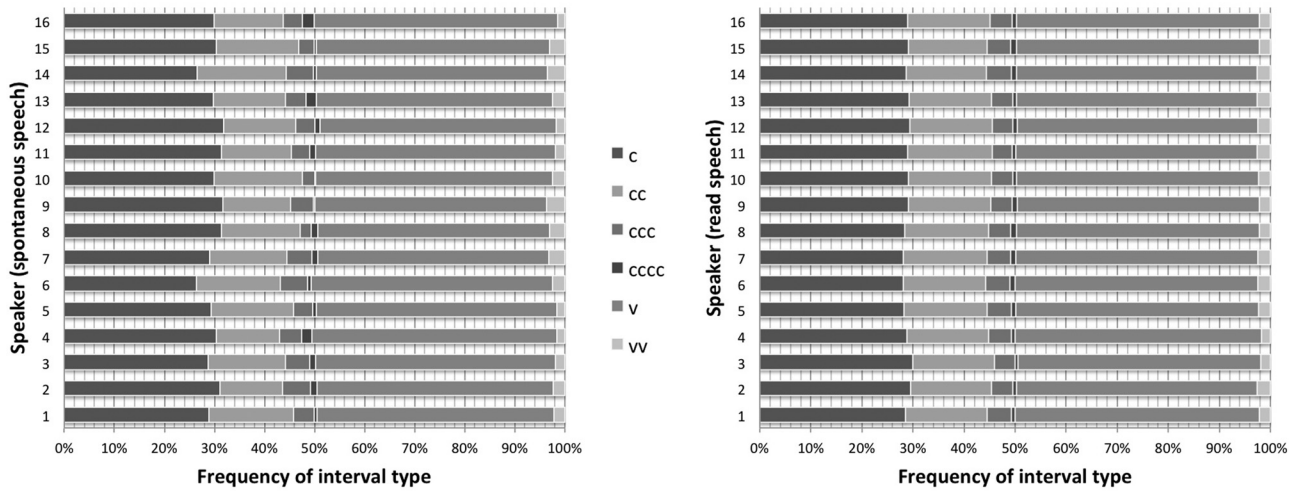

FIG. 3. Stacked bar-chart showing the relative frequency of interval types for each speaker (rows) in spontaneous (left, 16 sentences per speaker) and read (right, 256 sentences per speaker.

sentence origin. The main effect for speaker was significant for any rhythm measure for the raw data as well as the *z*-score data.

4. Normalizing the influence of sentence

All effects of between speaker-rhythmic variability obtained in Leemann *et al.* (2014) were replicated for the present measurement procedures using one-way ANOVAs with the factor speaker and repeated measures on speaker was calculated for each rhythm measure [R-code: aov (dependent ~ speaker) + error(sentence/speaker)]; $N = 4096$; 256 sentences × 16 speakers, α: 0.01: rate*CV*: $F[15,3825] = 477.6$, $p < 0.001$; *%V*: 106.4, < 0.001; Δ*V*(ln): 71.23, < 0.001; Δ*C*(ln): 36.7, < 0.001; Δ*peak*(ln): 31.28; <0.001. To test how many between-speaker comparisons were significant we carried out *post hoc* comparisons between the speakers using pairwise *t*-tests between each speaker pair (120 possible comparisons; *R*-function: pairwise.tst with Bonferroni correction). The number of significant *post hoc* comparisons (*R*-function: pairwise.tst) were as follows (absolute number of significant pairwise *t*-tests in brackets preceded by percentage from the total number of possible tests): rate*CV*: 82.5% (99/120), *%V*: 48.33% (58/120), Δ*V*(ln): 37.5% (45/120), Δ*C*(ln): 37.5% (45/120), Δ*peak*(ln): 37.5% (45/120). For this sentence set, the proportional number of significant between-speaker contrasts was much higher than in experiment 1, which indicates that larger datasets show clearer effects on between-speaker rhythmic differences. In particular rate*CV* shows much larger values compared to experiment 1 (it is possible that the extreme rate differences introduced in experiment 1 cancelled out the between-speaker rate variability to a high degree). To test whether a normalization for sentence variability could influence the number of significant between-speaker comparisons, we carried out the same pairwise *t*-tests between speakers (above) on the *z*-score transformed data and obtained the following results: rate*CV*: 86.67% (104/120), *%V*: 75.83% (91/120), Δ*V*(ln): 50.83% (61/120), Δ*C*(ln): 65% (78/120), Δ*peak*(ln): 52.5% (63/120). The number of significant comparisons increased notably [increase in percent-points: rate*CV*: 4.17, *%V*: 27.5, Δ*V*(ln): 13.33, Δ*C*(ln): 27.5, Δ*peak*(ln): 15]. rate*CV* was already close to a ceiling level (with over 80% of comparisons significant), which is why it was difficult to gain a large number of additional significant comparisons. Measures that showed lower numbers of significant pairwise comparisons based on the raw data increased these numbers drastically when normalized [Δ*V*(ln) and Δ*peak*(ln)]. In summary, between-speaker differences in rhythm were stronger when *z*-score normalization for the sentence was applied.

In Leemann *et al.* (2014) we tested the effects of speaking style for equal sentences by reducing the dataset to the 16 spontaneous sentences and the 16 read peers of each speaker ($N = 512$; 16 speakers × 16 sentences × 2 speaking

TABLE V. *F*-values and corresponding significance probabilities (*p*-values) for five (one for each rhythm measure) two-factor mixed design ANOVAs (speaker × sentence origin) with repeated measures on sentence origin ($N = 496$, $\alpha = 0.005$).

	Speaker (DOF: 15, 209)		Sentence origin (DOF: 1, 209)		Speaker: sentence origin (DOF: 15, 209)	
Dependent	Raw	*z*-score	raw	*z*-score	raw	*z*-score
rate*CV*	32.28 (**<0.001**)	38.55 (**<0.001**)	1.48 (0.226)	2.75 (0.099)	0.83 (0.647)	0.67 (0.812)
%V	7.48 (**<0.001**)	7.36 (**<0.001**)	3.11 (0.079)	2.45 (0.119)	0.76 (0.723)	0.66 (0.821)
Δ*V*(ln)	4.67 (**<0.001**)	4.8 (**<0.001**)	4.7 (0.031)	4.3 (0.039)	1.33 (0.187)	1.4 (0.149)
Δ*C*(ln)	4.02 (**<0.001**)	3.96 (**<0.001**)	0.04 (0.844)	0 (1)	0.41 (0.974)	0.57 (0.896)
Δ*peak*(ln)	3.17 (**<0.001**)	2.91 (**<0.001**)	0.76 (0.384)	0.26 (0.608)	0.54 (0.914)	0.56 (0.902)

TABLE VI. F-values and corresponding p-values (in brackets) for a two-factor mixed design ANOVA with repeated measures on speaking style (for raw and z-score data). Significant effects are highlighted in bold ($N = 512$; $\alpha = 0.01$).

	Factor speaker (between-subjects, DOF: 15, 240)		Factor speaking style (within-subjects, DOF: 1, 240)		Interaction speaker: Speaking style (DOF: 15, 240)	
Dependent	raw	z-score	raw	z-score	raw	z-score
rateCV	**11.53 (<0.001)**	**28.5 (<0.001)**	**12.84 (<0.001)**	**12.21 (0.001)**	**5.35 (<0.001)**	**5.98 (<0.001)**
$\%V$	**2.73 (0.001)**	**11.25 (<0.001)**	0.1 (0.751)	0.04 (0.838)	4.2 **(<0.001)**	**4.4 (<0.001)**
ΔV(ln)	**2.48 (0.002)**	**4.47 (<0.001)**	**9.34 (0.003)**	**9.95 (0.002)**	**3.77 (<0.001)**	**3.82 (<0.001)**
ΔC(ln)	1.54 (0.092)	**2.86 (<0.001)**	1.55 (0.214)	3 (0.085)	0.98 (0.476)	1.09 (0.366)
$\Delta peak$(ln)	1.91 (0.023)	**2.96 (<0.001)**	0.71 (0.4)	0.88 (0.35)	2.65 **(0.001)**	**2.43 (0.003)**

styles). Here we replicated these results for the rhythm measures used in the present study and we further applied the z-score normalization by sentence to test whether we can enhance the effects. For each rhythm measure we ran a two-way mixed design ANOVA with repeated measures on speaking style [R-code: aov(dependent ∼ speaker ∗ speaking style + error(sentence/speaking style))] on the raw as well as on the z-score data (Table VI). We found that all effects were highly significant for the factor speaker in the z-score data, which was not always the case for the raw data. We take this as evidence that normalization for sentence variability using z-scores is essential to obtain robust speaker-specific results. For ΔC(ln) the effect of speaker is significant (in the z-score data) but there is no speaking style effect and no interaction. As we received highly significant interactions in the case of all other rhythm measures we studied simple effects of speaking style and speakers to interpret the main effects. Table VII shows that we received highly significant effects of speaker in spontaneous as well as read speech for each rhythm measure for the z-score data. For the raw data there is no effect for ΔV(ln) and $\Delta peak$(ln) in spontaneously produced speech. So for some measures speaker-specific effects in spontaneous speech can only be obtained when the data is normalized for sentence variability. Simple effects of speaking style for each rhythm measure were calculated (not presented) with factor speaker for each of the two speaking style levels, either based on raw or z-score normalized data. α was 0.0006 (0.01/16; Bonferroni corrected for speaker). Results revealed that none of the tests was significant, neither for the raw nor for the normalized data which is further support for the lack of rhythmic variability between speaking styles.

TABLE VII. Simple effects of speaking style. F-values (DOF: 1, 255) with corresponding p-values (in brackets) for one-way ANOVAs for each temporal measure (rows) with factor speaker for each of the two speaking style levels, either based on raw or z-score normalized durations. Significant effects are highlighted in bold ($N = 256$; $\alpha = 0.005$).

	Spontaneous speech		Read speech	
Dependent	Raw	z-score	Raw	z-score
rateCV	6.96 **(<0.001)**	8.98 **(<0.001)**	16.65 **(<0.001)**	39.44 **(<0.001)**
$\%V$	3.21 **(<0.001)**	8.04 **(<0.001)**	2.51 **(<0.001)**	7.5 **(<0.001)**
ΔV(ln)	1.6 (0.07)	2.89 **(<0.001)**	4.37 **(<0.001)**	6.47 **(<0.001)**
$\Delta peak$(ln)	1.63 (0.07)	2.17 (0.01)	2.79 **(<0.001)**	3.35 **(<0.001)**

D. Discussion

Experiment 2 provided evidence that sentences vary in the complexity of their C- and V-intervals and that this variability has an influence on rhythmic measures to some degree. This result is in line with Prieto *et al*. (2012) who also found syllabic complexity to have an effect on rhythm scores within a language. The experiment further provided evidence that when speakers have the free choice of words and grammatical structures like in spontaneous speech, they vary to some degree in the structural complexity of C- and V-intervals. However, when we compared the rhythm scores of sentences that speakers constructed themselves with the scores for sentences that originated by other speakers we did not find any evidence that these phonotactic complexity differences could explain any of the between-speaker variability.

What might be the reason for between-speaker differences in speech rhythm in this experiment? The idiosyncratic choice of lexical and morphosyntactic structures (linguistic factors) did not reveal any difference. A possible explanation could again be an individual realization of prosodic features by speakers (e.g., speaker-idiosyncratic stress patterns) that might lead to a higher or lower variability of speech rhythm. Given, however, that the prosodic variability introduced by speaking style did not have an influence on between-speaker differences, we take this as further evidence that speaker-specific speech rhythm is not dependent on idiosyncratic prosody. To conclude, the results of experiment 2, like those of experiment 1, provided evidence for the hypothesis that the driving factor in between-speaker rhythmic variability are idiosyncrasies in the movements of the articulators.

V. GENERAL DISCUSSION

Both data from Standard German (experiment 1) and from Zurich Swiss German (experiment 2) revealed that there are strong differences between speakers in acoustically measurable speech rhythm even when prosodic and linguistic variability within speakers is strong. In both experiments we found strong effects of speaker and sentence but little to no influence of prosodic variability on speaker-specific results. Experiment 2 showed clearly, linguistic structural characteristics of a speaker were not responsible for idiosyncratic rhythm. Given the three possible factors that might drive speaker-specific rhythm (see the Introduction), it now

seems feasible to put prosodic and linguistic influences into the background. It is thus increasingly likely that individual ways of operating the articulators should influence speaker-specific temporal variability.

How could a speaker-specific way of moving the articulators result in individual patterns in the measures we have chosen for the present study? Two types of measures were present, a temporal vocalic-consonantal ratio measure (*%V*) and three measures of durational variability, vocalic [ΔV(ln)], consonantal [ΔC(ln)], and inter-syllable amplitude peak variability [$\Delta peak$(ln)]. Individuality in movements of the articulators can be either acquired or it can be a result of the genetically determined dimensions of the articulators (see Sec. I) and possibly it is a complex mixture of both factors. One conceivable assumption might be that some vowel-consonant transitions underlying certain movements are more affected than others. A movement that requires the tongue to reach from the front to the back (as in /θu:/) or the jaw to move from a closed to an open position (e.g., /ma:/) might be more affected than movements where not much articulatory change is involved (e.g., /ku:/). Accordingly, it should be the case that individual vowels and consonants do not equally contribute to the vocalic and consonantal variability we obtained in our results. Therefore, more refined measures which focus on particular consonants, vowels and consonant-vowel transitions may lead to clearer between-speaker results. If speaker-specific temporal characteristics were stronger for /θu:/ as opposed to /ku:/ type syllables this should be further evidence for an articulatory explanation of between-speaker rhythmic variability. It would be interesting to refine such measures and test these hypotheses.

We included rhythm measures based on consonantal and vocalic interval durations [*%V*, ΔV(ln), ΔC(ln)] as well as a measure based on the amplitude envelope [$\Delta peak$(ln)] in our study. Both types of measures showed rather similar results with the exception that in experiment 1, the interval measures showed more consistent results for between-speaker variability (in terms of descriptive magnitude of the effects, Fig. 1, and in terms of the number of significant *post hoc* comparisons). In experiment 2 there were no such differences (in particular, for the between-speaker comparison based on read speech). In general we can conclude that both durational characteristics of speech intervals as well as the speech amplitude envelope vary between speakers and sentences.

What do the results tell us about language-specific rhythmic variability? The results we obtained do not stand in contrast with previous results on language-specific rhythmic variability. They might rather explain why some studies obtained inconsistent results for between-language variability (Grabe and Low, 2002; Arvaniti, 2012). Since different languages are characterized, in particular, by different phonotactic and phonological phenomena influencing consonantal and vocalic durations, it seems conceivable that language variability exists in addition to within-language speaker and sentence variability. The results from the present study, however, underline the point by Wiget *et al.* (2010) that only studies using large numbers of speakers and sentences can lead to representative between-language results.

What do the measures applied tell us about speech rhythm? This question is difficult to answer since there is no unified and generally accepted definition of speech rhythm (see Sec. I). Early theories of speech rhythm which were mainly concerned with between-language rhythmic variability emphasized auditory phenomena, claiming that some languages sound rhythmically differently from others (Ramus *et al.*, 1999). In more recent discussions on speech rhythm and its acoustic correlates, such auditory characteristics seem to have played a secondary role (Grabe and Low, 2002; Dellwo, 2006; Arvaniti, 2012). If speech rhythm is about auditory characteristics of speech, then we may expect that variability between strongly varying prosodic realizations of utterances should affect such auditory characteristics in some way. Since prosodic changes had little effect on rhythm measures in our study we take this as evidence that the acoustically measurable rhythmic stability we obtained between prosodically varying utterances probably does not reflect auditory rhythmic characteristics of the signal well. So it might be more appropriate to refer to such measures as suprasegmental-timing rather than rhythm measures. What is quite surprising in this respect is that both the measures based on consonantal and vocalic interval durations and the measure based on the amplitude envelope of speech show very similar results. An explanation for this might be that also the temporal characteristics of the amplitude envelope are not as salient in terms of auditory speech rhythm as previously assumed (Tilsen and Arvaniti, 2013). An alternative explanation, however, is that the temporal anchor points that we chose (syllabic amplitude peak points) are not strong correlates of perceptual rhythmic beats in the signal (see Sec. I C). Given that we obtained strong results for between-speaker effects and under the assumption that articulatory factors are the driving source for this variability, we take this as evidence for our hypothesis that the amplitude peak points may reflect important speaker-specific movement characteristics. Since amplitude peak points are much easier to extract automatically than consonantal and vocalic intervals, they may be more applicable for automatic systems.

What implication do the results have for our human ability to identify speakers based on their voices? From the field of between-language rhythmic characteristics there is strong behavioral evidence that human listeners perceive differences between languages based on the type of durational variability examined in the present study. Experiments have shown that adult human listeners (Ramus and Mehler, 1999), as well as newborns (Nazzi *et al.*, 1998; Ramus, 2002) can distinguish between languages from different rhythmic classes. This lead to the argument that such rhythmic characteristics are acquired at a pre-linguistic stage and that they might aid listeners (e.g., infants growing up in a bilingual environment) segregate between different languages (Ramus *et al.*, 1999). Since durational characteristics of consonantal and vocalic intervals are perceptually salient between languages it seems conceivable that between-speaker variability is salient too. It would be interesting to test this hypothesis further in behavioral experiments.

What applications could between-speaker rhythmic variability have? The results presented in the present research

might be relevant for any type of application where speaker-specific information plays a role, e.g., technical speaker identification and forensic phonetic speaker comparison (Leemann *et al.*, 2014). For such applications, we argue that, in particular, our approach of maximizing between-speaker differences by normalizing for sentence variability using *z*-scores is an important finding. However, there might yet be another feature making rhythm measures appropriate for speaker identification purposes. Speaker identification applications make strong use of frequency domain variables like fundamental and formant frequencies or the entire spectral envelope characteristics because they are shaped by individual anatomic characteristics of the vocal tract. These variables, however, are highly claimed by other channels for the transfer of functional linguistic and paralinguistic information. While the speech signal is highly organized in time it seems that the suprasegmental temporal organization is not used in an elaborate way to convey linguistic or paralinguistic information. In the cases in which speakers use variables to form functional contrasts in speech, they need active control over these parameters to modulate them and their perceptual system must be tuned in on them. In other cases such a control might not be necessary to the same degree. Speakers might thus be much less capable of controlling rhythmic parameters than they are of controlling intonation or stress, for example (at least for the languages of which we know that rhythm is not a primary carrier of linguistic information). This might be particularly relevant for identification purposes in which speakers are non-cooperative (forensic phonetic speaker comparison) and frequently apply voice disguise techniques to impede on identification.

VI. CONCLUSIONS

We have shown that rhythm measures based on consonantal and vocalic interval durations as well as temporal characteristics of the amplitude envelope vary strongly between speakers while within-speaker prosodic and linguistic variability has little effect on them. It seems more likely that speaker individual control mechanisms of the articulators are responsible for the obtained between speaker differences. It would be interesting to test this hypothesis with articulatory measures in the future. Further research is also necessary to address the perceptual salience of rhythmic temporal characteristics for auditory speaker identification.

ACKNOWLEDGMENTS

This research was funded by Grant No. 100015_135287 of the Swiss National Science Foundation and by Grant No. GRS-027/13 by the Gebert Rüf Foundation. The authors wish to thank Stephan Schmid for helpful comments on draft versions of the paper; Paul Iverson, Martin Meyer, and his lab members at Zurich University; and Sandra Schwab and Mattia Molinaro for helpful comments on statistical procedures.

APPENDIX A: BonnTempo CORPUS

The reading text of the BonnTempo corpus. The parts between the vertical lines are syntactic intervals for which the temporal measures were calculated in the corresponding part of the acoustic signal (seven units in total).

| *Am nächsten Tag fuhr ich nach Husum.* | *Es ist eine Fahrt ans Ende der Welt.* | *Hinter Giessen werden die Berge und Wälder eintönig,* | *hinter Kassel die Städte ärmlich* | *und bei Salzgitter wird das Land flach und öde.* | *Wenn bei uns Dissidenten verbannt würden,* | *würden Sie ans Steinhuder Meer verbannt.* |

APPENDIX B: TEVOID CORPUS

The first 20 of the 256 sentences of the TEVOID Corpus.

(1) *So s Typische was sich d Lüüt vorschteled isch Kurator.*
(2) *Ich han Freiziit.*
(3) *Ich han käi äigeni Band.*
(4) *Ich bin wäge Spraachwüsseschaft dänn usegheit.*
(5) *Das han i cool gfunde.*
(6) *Mini Mueter isch ä no nie z Wien gsi.*
(7) *Dänn mues ich ä no überlegge, was mis nöie Hauptfach wird.*
(8) *Ich ha jetz äifach vergliichendi Spraachwüsseschafte gno.*
(9) *Ich ha mich ä nie würklich beworbe.*
(10) *Wänn ich halt im Usland wär, hett ich das zmindescht mal für es Semeschter nöd.*
(11) *Chasch ja nöd nöime andersch go studiere mit Erasmus.*
(12) *Si liit det am Bode.*
(13) *Zwar isch das Ganze im ne fiktive Königriich.*
(14) *Ich wäis nöd werum si so abglänkt isch.*
(15) *Säge mer emaal ich fahr uf Oerlike.*
(16) *Mit em Zug sälber zwänzg Minute.*
(17) *Wil's äifach di zwäiti Spraach isch.*
(18) *Das git's äigentlich i käinere andere Spraach.*
(19) *Tschechisch han i käi Phonetik ghaa.*
(20) *Ich glaub mi händ so chli ä drüber gredt ghaa.*

[1]See Sec. III B 3 for a definition of *C*- and *V*-intervals. Note that *C*- and *V*-intervals can contain one or more consonantal and vocalic segments respectively (henceforth: c- and v-segments).

Arvaniti, A. (**2012**). "The usefulness of metrics in the quantification of speech rhythm," J. Phonetics **40**, 351–373.

Boersma, P., and Weenink, D. (**2013**). "Praat: doing phonetics by computer" [Computer program]. Version 5.3.51, http://www.praat.org (Last viewed 4 February 2015).

Caspers, J., and van Heuven, V. J. (**1995**). "Effects of time pressure on the choice of accent-lending and boundary-marking pitch configuration in Dutch," *Proceedings of Eurospeech 2*, pp. 1001–1004.

Dauer, R. M. (**1983**). "Stress-timing and syllable-timing reanalyzed," J. Phonetics **11**, 51–69.

Dellwo, V. (**2006**). "Rhythm and speech rate: A variation coefficient for deltaC," in *Language and Language-Processing*, edited by P. Karnowski and I. Szigeti (Peter Lang, Frankfurt am Main), pp. 231–241.

Dellwo, V. (**2009**). "Choosing the right rate normalization methods for measurements of speech rhythm," in *Proceedings of AISV*, pp. 13–32.

Dellwo, V. (**2010**). "Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and

perceptual evidence," Ph.D. dissertation, Universität Bonn, Bonn, Germany, pp. 1–185.
Dellwo, V. (**2013**). http://www.pholab.uzh.ch/leute/dellwo/software.html (Last viewed 4 January 2015).
Dellwo, V., and Fourcin, A. (**2013**). "Rhythmic characteristics of voice," in *Travaux Neuchâtelois de Linguistique* (TRANEL), pp. 87–107.
Dellwo, V., Huckvale, M., and Ashby, M. (**2007**). "How is individuality expressed in voice? An introduction to speech production and description for speaker classification," in *Speaker Classification I*, edited by C. Müller (Springer, Berlin), pp. 1–20.
Dellwo, V., Leemann, A., and Kolly, M. J. (**2012**). "Speaker idiosyncratic rhythmic features in the speech signal," in *Proceedings of Interspeech*, pp. 1584–1587.
Dellwo, V., Steiner, I., Aschenberner, B., Dankovicova, J., and Wagner, P. (**2004**). "The BonnTempo-Corpus and BonnTempo-Tools: A database for the study of speech rhythm and rate," in *Proceedings of Interspeech ICSLP*, Jeju Island/Korea, pp. 777–780.
Dellwo, V., and Wagner, P. (**2003**). "Relations between language rhythm and speech rate," in *Proceedings of the International Congress of Phonetics Science*, pp. 471–474.
Fougeron, C., and Jun, S. A. (**1998**). "Rate effects on French intonation: Prosodic organization and phonetic realization," J. Phonetics **26**, 45–69.
Grabe, E., and Low, E. L. (**2002**). "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology C*, edited by N. Warner Gussenhoven (Mouton de Gruyter, Berlin), Vol. 7, pp. 515–545.
Howell, P. (**1988**). "Prediction of P-center location from the distribution of energy in the amplitude envelope: I," Percept. Psychophys. **43**, 90–93.
Howell, P., and Kadi-Hanifi, K. (**1991**). "Comparison of prosodic properties between read and spontaneous speech material," Speech Commun. **10**, 163–169.
Johnson, C., and Hollien, H. (**1984**). "Speaker indentification [sic!] utilizing selected temporal speech features," J. Phonetics **12**, 319–326.
Kohler, K. J. (**1983**). "F0 in speech timing," Arbeitsberichte Institut für Phonetik, Universität Kiel, Vol. 20, pp. 57–97.
Leemann, A., Kolly, M.-J., and Dellwo, V. (**2014**). "Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison," Forensic Sci. Int. **238**, 59–67.
Lieberman, P., Katz, W., Jongman, A., Zimmerman, R., and Miller, M. (**1985**). "Measures of the sentence intonation of read and spontaneous speech in American English," J. Acoust. Soc. Am. **77**, 649–657.
Loukina, A., Kochanski, G., Rosner, B., Keane, E., and Shih, C. (**2011**). "Rhythm measures and dimensions of durational variation in speech," J. Acoust. Soc. Am. **129**, 3258–3270.
Loula, F., Prasad, S., Kent, H., and Shiffrar, M. (**2005**). "Recognizing people from their movement," J. Exp. Psychol.: Human Percept. Perform. **31**, 210–220.
McDougall, K. (**2004**). "Speaker-specific formant dynamics: An experiment on Australian English /aI/," Int. J. Speech Lang. Law **11**, 103–130.
McDougall, K. (**2006**). "Dynamic features of speech and the characterisation of speakers: Towards a new approach using formant frequencies," Int. J. Speech Lang. Law **13**, 89–126.
Mendoza, E., Carballo, G., Cruz, A., Fresneda, M. D., Muñoz, J., and Marrero, V. (**2003**). "Temporal variability in speech segments of Spanish: Context and speaker related differences," Speech Commun. **40**, 431–447.
Morton, J., Marcus, S., and Frankish, C. (**1976**). "Perceptual centers (P-centers)," Psychol. Rev. **83**, 405–408.
Nazzi, T., Bertoncini, J., and Mehler, J. (**1998**). "Language discrimination by newborns: Toward an understanding of the role of rhythm," Exp. Psychol. **24**, 756–766.
Nixon, M. S. (**2008**). "Automated human recognition by gait using neural network," in *First Workshops on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6.
Nolan, F. (**2002**). "Intonation in speaker identification: An experiment on pitch alignment features," Forensic Ling. **9**, 1–21.
Nolan, F. (**2009**). *The Phonetic Bases of Speaker Recognition* (Cambridge University Press, Cambridge), pp. 1–232.
O'Dell, M., and Nieminen, T. (**1999**). "Coupled oscillator model of speech rhythm," in *Proceedings of the 14th ICPhS*, pp. 1075–1078.
O'Shaughnessy, D. (**1984**). "A multispeaker analysis of durations in French paragraphs," J. Acoust. Soc. Am. **76**, 1664–1672.
Perrier, P. (**2012**). "Gesture planning integrating knowledge of the motor plant's dynamic: A literature review from motor control and speech motor control," in *Speech Planning and Dynamics*, edited by S. Fuchs, M. Weirich, D. Pape, and P. Perrier (Peter Lang, Frankfurt am Main), pp. 191–238.
Prieto, P., del Mar Vanrell, M., Astruc, L., Payne, E., and Post, B. (**2012**). "Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish," Speech Commun. **54**, 681–702.
Ramus, F. (**2002**). "Language discrimination by newborns," Ann. Rev. Lang. Acquis. **2**, 85–115.
Ramus, F., and Mehler, J. (**1999**). "Language identification based on suprasegmental cues: A study based on resynthesis," J. Acoust. Soc. Am. **105**, 512–521.
Ramus, F., Nespor, M., and Mehler, J. (**1999**). "Correlates of linguistic rhythm in the speech signal," Cognition **73**, 265–292.
Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., and Stolcke, A. (**2005**). "Modelling prosodic feature sequences for speaker recognition," Speech Commun. **46**, 455–472.
Tilsen, S., and Arvaniti, A. (**2013**). "Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages," J. Acoust. Soc. Am. **134**, 628–639.
Tilsen, S., and Johnson, K. (**2008**). "Low-frequency Fourier analysis of speech rhythm," J. Acoust Soc. Am. **124**, EL34–EL39.
Trouvain, J., and Grice, M. (**1999**). "The effect of tempo on prosodic structure," in *Proceedings of the ICPhS*, pp. 1067–1070.
van den Heuvel, H., Rietveld, T., and Cranen, B. (**1994**). "Methodological aspects of segment- and speaker-related variability. A study of segmental durations in Dutch," J. Phonetics **22**, 389–406.
Vaissière, J. (**1983**). "Language-independent prosodic features," in *Prosody: Models and Measurement*, edited by A. Cutler and D. R. Ladd (Springer, New York), pp. 53–66.
White, L., and Mattys, S. L. (**2007**). "Calibrating rhythm: First language and second language studies," J. Phonetics **35**, 501–522.
Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., and Mattys, S. L. (**2010**). "How stable are acoustic metrics of contrastive speech rhythm?," J. Acoust. Soc. Am. **127**, 1559–1569.
Yoon, T. J. (**2010**). "Capturing inter-speaker invariance using statistical measures of speech rhythm," in *Proceedings of Speech Prosody*, Chicago, IL, Vol. 5, pp. 1–4.